DOCUMENT RESUME

ED 418 999 TM 028 294

AUTHOR Li, Yuan H.; Lissitz, Robert W.

TITLE An Evaluation of Multidimensional IRT Equating Methods by Assessing the Accuracy of Transforming Parameters onto a Target Test Metric.

PUB DATE 1998-04-00

NOTE 48p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Diego, CA, April 12-16, 1998).

PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS *Equated Scores; Estimation (Mathematics); *Item Response Theory

IDENTIFIERS Item Parameters; *Multidimensional Approach

ABSTRACT

The metric of the multidimensional item response theory (MIRT) item parameter estimates is usually referred to as reference axes that are orthogonal and of unit length. This is due to the fact that most MIRT parameter estimation programs solve the identification problem by requiring that multidimensional abilities be distributed as multivariate normal distribution, N (0, I). Under this circumstance, the equated group's reference system can be transformed into the base group's reference system by a composite transformation: an orthogonal procrustes rotation, a translation transformation, and a single dilation. Based on this composite transformation, three sets of MIRT equating methods have been developed and evaluated in this study. The results indicate that the best MIRT equating method is an unbiased, effective, and consistent estimator producing accurate transformation parameter estimates when the errors in the estimation of item parameters were purposely manipulated. In addition, this MIRT equating method is capable of successfully recovering parameters, especially for the item parameters, under well-fitting model conditions. (Contains 17 figures, 4 tables, and 35 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 418 999

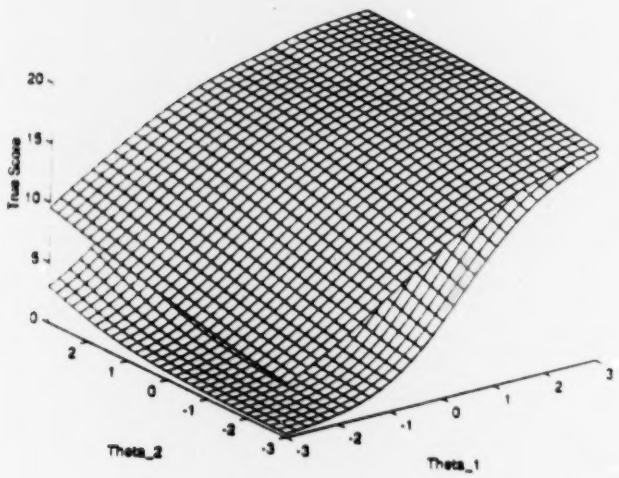
An Evaluation of Multidimensional IRT Equating Methods by Assessing the Accuracy of Transforming Parameters onto a Target Test Metric

Yuan H. Li*

University of Maryland & Prince George's County Public Schools

Robert W. Lissitz*

University of Maryland at College Park



PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Yuan H Li

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Paper presented at the annual meeting of the National Council on Measurement in Education
April 14-16, 1998, San Diego, California

** The authors would like to convey special thanks to Mark Reckase, Rafael De Ayala, C. M. Dayton and Frank Alt for their insightful suggestions on this paper*

TM028294

BEST COPY AVAILABLE

An Evaluation of Multidimensional IRT Equating Methods by Assessing the Accuracy of Transforming Parameters onto a Target Test Metric

Abstract

The metric of the MIRT (multidimensional item response theory) item parameter estimates is usually referred to reference axes that are orthogonal and of unit length due to the fact that most MIRT parameter estimation programs solve the identification problem by requiring that multidimensional abilities be distributed as multivariate normal distribution, N(**0**, **I**). Under this circumstance, the equated group's reference system can be transformed into the base group's reference system by a composite transformation: an orthogonal procrustes rotation, a translation transformation and a single dilation. Based on this composite transformation, three sets of MIRT equating methods have been developed and evaluated in this study. The results from this study indicate that the best MIRT equating method is an unbiased, effective and consistent estimator producing accurate transformation-parameter estimates when the errors in the estimation of item parameters were purposely manipulated. In addition, this MIRT equating method is capable of successfully recovering parameters, especially for the item parameters, under well-fitting model conditions.

Index Terms: Item Response Theory (IRT); Test Equating; Item Linking; Multidimensional Item Response Theory (MIRT); Simulation Study.

I. Introduction

A. Motivation

Item response theory (IRT) consists of a family of probabilistic models that hypothesize the relationship between an examinee's latent ability and a correct response to an item. It is often assumed that only one latent trait is necessary to account for variations in examinee's item response vectors (Hambleton & Swaminathan, 1985). However, the test-examinee interaction process can be quite complex since a set of test items may be sensitive to several traits; and a group of examinees may vary in several latent abilities (Ackerman, 1992). It becomes apparent that practitioners may encounter problems applying unidimensional IRT to multidimensional datasets. These problems may jeopardize the invariant feature of the unidimensional IRT models (Ackerman, 1992; McKinley & Mills, 1985).

Test equating methods have been developed to adjust for differences in item characteristics such as item difficulties and discriminations among test forms so that examinees' scores reported from different test forms can be converted into each other (Kolen & Brennan, 1995). When two tests are equated, it is of course desirable to try to ensure that the two test datasets are unidimensional. Unfortunately, this condition may not be met in most testing situations. Consequently, inaccurate test score reports could occur when reporting scales from different tests, which are equated using unidimensional IRT equating when latent dimensions on each test dataset are multidimensional. Researchers who have used multidimensional IRT (MIRT) analysis (e.g. Ackerman, 1992, 1994; Bock, Gibbons & Muraki, 1988; Reckase, 1985) have indicated that MIRT models more adequately explain both simulated and real multidimensional data than do unidimensional models. In other words, the feature of invariant parameters of MIRT can be retained. From the perspective of MIRT equating, test practitioners can ask if two test datasets are essentially multidimensional, and can MIRT equating produce correct test scores. That is the central issue of MIRT equating.

B. Background of MIRT Equating Methods

When the same set of test items is administered to two groups (called the base group and the equated group) with differing ability profiles and two sets of test response data are calibrated separately, two sets of different numerical values of parameter estimates are obtained. The reason given is that the origins and units of the two reference systems (ability-dimensions) are defined independently and the reference systems may be rotated (Reckase, 1985). The MIRT equating method is used for transforming the equated group's reference system into the base group's reference system by re-rotating its reference system, re-scaling its unit length and shifting its point of origin.

The metric of the MIRT item parameter estimates is usually referred to reference axes that are orthogonal and of unit length due to the fact that most MIRT parameter estimation programs solve the identification problem (or result in a unique solution) by requiring multidimensional abilities ($\underline{\theta}$) be distributed as multivariate normal distribution, N($\mathbf{0}$, $\mathbf{I}$) (Mislevy, 1986).Under this circumstance, the equated group's reference system can be transformed into the base group's reference system by a composite (Schonemann & Carroll, 1970): an orthogonal procrustes rotation, a translation transformation, and a single dilation (or contraction). A single dilation parameter is preferred because of two main reasons: (1) it provides a more tractable mathematical problem and (2) in most cases the variance across dimensions will be similar enough that a single dilation parameter will provide reasonable accuracy. This kind of composite transformation takes into account the properties of the metric of MIRT item parameters defined by MIRT computer programs and holds the features of : (1) symmetry, that is, the sum of squared errors is the same whether the equated test is linked to the base test or the base test is linked to the equated test, and (2) retains original angles and relative positions of all pairs of item discrimination parameters in space while referring them to the base group's coordinate system.

In reviewing the research on MIRT equating methods, four particularly relevant studies (Hirsch, 1989; Oshima & Davey, 1994; Oshima, Davey & Lee, 1997; Thompson, Nering & Davey, 1997) were identified. The first MIRT equating study conducted by Hirsch (1989)

provided practitioners with valuable knowledge of the MIRT equating principles. The MIRT method employed in Hirsch's study was, however, rather complicated. For instance, in the two-dimensional MIRT case, four rotation matrixes, two dilation parameters and two location parameters need to be estimated. In contrast, the four MIRT equating methods developed in the Oshima et al.' study were much more straightforward (refer to Oshima et al., 1997). The main approach of each of the four MIRT equating methods introduced in the Oshima et al.' study is to simultaneously estimate a rotation matrix and a translation vector by minimizing a mathematical function such as defined by squared differences of two test characteristic surfaces. However, it should be pointed out that no dilation parameter was found or defined in the Oshima et al's MIRT equating methods. The rotation matrix produced by Oshima et al's study can be formed by a variety of composite transformations, for instance, a scalar times a nonorthogonal transformation matrix.

On the other hand, Thompson et al (1997) attempted to develop a MIRT linking method that can be employed to equate tests without common items or common examinees under the circumstance that different tests are randomly assigned to a very large number of examinees. They "assume that the origin, axes, and correlations between axes are the same between groups since the groups are randomly equivalent(p 5)." However, the arbitrary rotation for each group's reference system occurs during the process of parameter estimation as it happens in factor analysis. This phenomena is known as rotational indeterminacy which is tackled by identifying similar item content "clusters" (note: not items) on different test forms and then by rotating them in the same multidimensional-reference system. The approach taken to identify item clusters on different test forms can be found in Reckase, Thompson and Nering (1997). Whether this linking procedure can be successfully accomplished relies on the accuracy of identifying similar item content clusters on multiple test forms. Consequently, this MIRT equating approach is still experimental and more evidence to clarify those uncertain issues, as pointed out in the Thompson et al's study (1997), is still needed.

BEST COPY AVAILABLE

C. Statements of Research Question

The accuracy of parameter estimates can be assessed by using BIAS (average differences between the true parameters and the corresponding estimates) and RMSE (root mean square errors) (refer to Skaggs & Lissitz, 1988). From a practical perspective, the MIRT equating methods used in previous studies were not formally evaluated in terms of the BIAS and RMSE indices of the transformation-parameter estimates, or the transformed item and ability estimates. Apparently, more study is needed to evaluate these techniques. In addition, developing more MIRT equating techniques, especially those that can maintain such features as the composite transformation (Schonemann & Carroll, 1970), is critical. Accordingly, based on the above composite transformation, three sets of MIRT equating methods have been developed in this study and have been evaluated by using the BIAS and RMSE criterion. The three MIRT equating methods are illustrated in the next section.

Sample size, the shape of examinees' abilities and the characteristic of test items can cause errors in the parameter estimates. A mathematical expression for this relationship has been developed by Thissen and Wainer (1892) for a family of IRT models and is used in this study for modeling error in the simulation of parameter estimates. The issue of which MIRT equating method among the three ones developed can produce the most accurate transformation of parameter estimates was evaluated under conditions in which errors in the estimation of item parameters were purposely manipulated. The issue of "Can this best MIRT equating method produce accurate and reliable results for linking item parameters onto a common scale?" was then examined under conditions in which the examinee's ability characteristics were manipulated.

It is usually assumed that a MIRT model should be applied to a multidimensional dataset. It could happen that practitioners will apply a MIRT model to a test dataset, in which only one latent dimension exists. Practitioners would then be concerned with whether a MIRT model can explain a unidimensional test dataset. Thus, a unidimensional dataset was created and an MIRT model will be applied to it. The issue of parameter recovery was explored.

BEST COPY AVAILABLE

II. MIRT Equating Methods for Dichotomous Item Response Data

A. Multidimensional Logistic IRT Model

The probability of a correct response, $u_{ij}=1$, by person j to item i, given an individual's m-dimensional latent abilities, $\underline{\theta}_j$, is (refer to Reckase, 1985):

$$P(u_{ij}=1 \mid \underline{a}_i, d_i, c_i, \underline{\theta}_j) = c_i + (1-c_i)\frac{e^{Z_{ij}}}{1+e^{Z_{ij}}} \qquad (1)$$

where,

$$Z_{ij} = \sum_{k=1}^{m} a_{ik}\theta_{jk} + d_i \qquad (2)$$

$\underline{a}_i$ is a m-dimensional vector of item discrimination parameters,

d_i is a location parameter related to item difficulty and

c_i is a pseudo-guessing parameter.

The above model is a multidimensional extension of the three-parameter logistic model (M3PL). Since the terms in Equation 2 are additive, being low on one latent trait can be compensated for by being high on the other latent traits. Thus, this model is called a compensatory model (Reckase, 1985). A multidimensional extension of the two-parameter logistic model (M2PL) is attained if the guessing parameter c_i is constrained to zero for all items in Equation 1 above.

B. Rotational Indeterminacy and Scale Indeterminacy

There are two main issues that need attention in the invariant MIRT parameters (Carlson, 1987; Hirsch, 1989). They are: Rotational Indeterminacy and Scale Indeterminacy.

Rotational Indeterminacy: A special feature of the MIRT model is the joint rotational indeterminacy of the vector of item discriminations, $\underline{a}_i$, and the vector of abilities, $\underline{\theta}_j$. The axes of the ability space can be rotated by pre-multiplication by a matrix $\mathbf{T}_R$ and the $\underline{a}'_i$ simultaneously is post-multiplied by the inverse of $\mathbf{T}_R$, $\mathbf{T}_R^{-1}$, so that the probability of a correct response given ability values is not altered (refer to Hirsch, 1989).

$$\underline{\theta}_j^R = \mathbf{T}_R \underline{\theta}_j \quad (3)$$

$$\underline{a}_i^R = \underline{a}_i \mathbf{T}_R^{-1} \quad (4)$$

Thus, an infinite pairwise number of θ_j^R and a_i^R parameter estimates retain the $P(u_{ij}=1 \mid \underline{a}_i, d_i, \underline{\theta}_j)$ unaltered. In practice, the MIRT parameter estimation program TESTFACT (Wilson, Wood & Gibbons, 1991) solves the identification problem by imposing multidimensional-θ to be distributed with iid $N(\mathbf{0}, \mathbf{I})$ at the end of the estimation process. In addition, the MIRT item discrimination estimates from any solution (e.g., the full-information item analysis, Muraki & Engelhard, 1985; Bock, Gibbons & Muraki, 1988) can be arbitrarily rotated by some criterion, for instance, "Varimax".

Scale Indeterminacy: Similar to the unidimensional models, the MIRT model is invariant to the origin shift and unit change. The following transformations are used to overcome scale indeterminacy in the parameter estimation by standardizing the ability estimates for each dimension to zero mean and unit variance after each step in which parameters are estimated (Carlson, 1987). In the two dimensional framework, the point of origin of each ability dimension (θ_{j1}, θ_{j2}) can be shifted by subtracting the corresponding mean ($\mu_{\theta 1}$, $\mu_{\theta 2}$) of the ability estimates on each dimension. Meanwhile, the unit of each ability dimension can be rescaled by dividing by the corresponding standard deviation ($\sigma_{\theta 1}$, $\sigma_{\theta 2}$) of the ability estimates (refer to Carlson, 1987; Hirsch, 1989).

$$\theta_{j1}^{*} = \frac{\theta_{j1} - \mu_{\theta 1}}{\sigma_{\theta 1}} \quad (5)$$

$$\theta_{j2}^{*} = \frac{\theta_{j2} - \mu_{\theta 2}}{\sigma_{\theta 2}} \quad (6)$$

The probability of a correct response for any item given an individual's ability values remains unchanged if item discriminations (a_1, a_2) are rescaled by multiplying by the corresponding standard deviations of the ability estimates.

$$a_{i1}^{\#} = \sigma_{\theta 1}\, a_{i1} \quad (7)$$

$$a_{i2}^{\#} = \sigma_{\theta 2}\, a_{i2} \quad (8)$$

and d is rescaled by:

$$d_i^{\#} = d_i + a_{i1}\mu_{\theta 1} + a_{i2}\mu_{\theta 2} \quad (9)$$

Through these transformations, the original ability estimates (θ_{j1} or θ_{j2}) on each dimension are converted into the z-scale scores ($\theta_{j1}^{\#}$ or $\theta_{j2}^{\#}$) having zero mean and unit variance. As seen in Equation 7 or 8, each transformed value of a-parameter ($a_{i1}^{\#}$ or $a_{i2}^{\#}$) on each dimension equal the product of the original a-parameter and the corresponding standard deviation (a constant, $\sigma_{\theta 1}$ or $\sigma_{\theta 2}$) of ability estimates. Thus, the new variance of the transformed-a-parameter on each dimension equals the product of the original variance and the corresponding variance of ability estimates.

Reckase (1997a) indicated that "If the correlations among the θ-dimensions are constrained to be 0.0, then the observed correlations among the item scores will be accounted for solely by the a-parameters (p275)." Similarly, if the variances of multidimensional abilities are constrained to be one, the a-parameter estimates produced from most MIRT estimation programs will take into account the original heterogeneous variances of multidimensional abilities. By imposing $\underline{\theta}$ to be distributed as N (**0**, **I**), the original (or unstandardized) heterogeneous variances and covariances of multidimensional abilities are captured by the variances and covariances of multidimensional a-parameter estimates.

C. Multidimensional IRT Equating

From the geometric perspective, the numerical estimates of the MIRT item parameters depend upon an arbitrary reference system (θ-dimensions). For instance, as illustrated in Figure 1, the basis vectors of the equated form are different from those of the base form. Those differences are (refer to Green, 1976):

(a). The axes for the base form are B_1 and B_2. The axes for the equated form are E_1 and E_2 which can be rotated into the base form's corresponding axes by premultiplying or postmultiplying an orthogonal rotation matrix, **T** (or $\mathbf{T}^{-1}$);

(b). The point of origin for the base form is O_B; the point of origin for the equated form is O_E, which is moved from O_B to O_E. The coordinates of point O_B and O_E are defined as (0,0) and (m_1, m_2), respectively, where m_1 is the length of the shift from the point of origin of the first dimension in the base form to the equated form and m_2 is the length of the shift for the second dimension. These m_1 and m_2 are translation coefficients that are used to translate the point of origin for the equated form into the point of origin for the base form; and

(c). The unit for the base form equals the segment of point O_B and point U_B and the unit for the equated form equals the segment of point O_E and point U_E. The ratio of the unit for the equated form to the unit in the base form represents the dilation (or contraction) coefficient, k which is used to dilate (or to contract) the unit for the equated form to the unit of the base form.

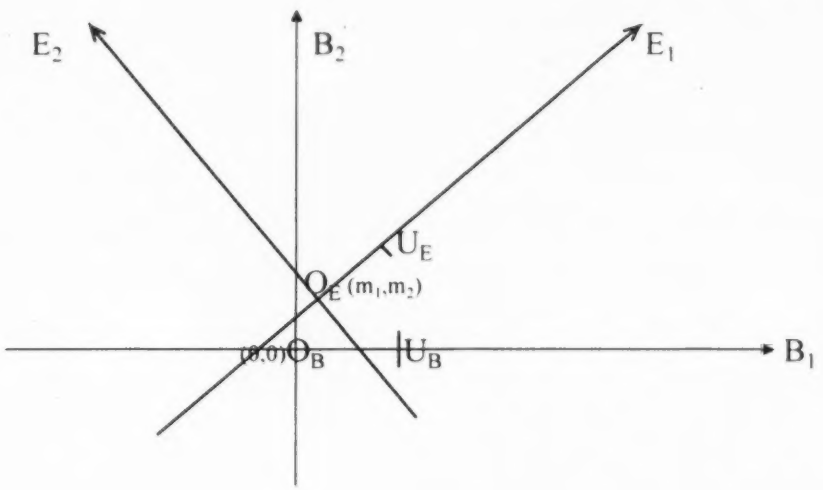


Figure 1. A Composite Transformation: A Rotation, a Translation and a Central Dilation.

The above composite transformation of the scale and the rotational transformations can be expressed using a matrix notation. That is, if the model selected fits the data (see McKinley & Mills, 1985; Tam & Li, 1997, for reviews of popular methods for evaluating model-data fit), $\underline{a}_{iE}' \underline{\theta}_{jE} + d_{iE}$ (refer to Equation2) remains unaltered when both item ($\underline{a}_{iE}'$, d_{iE}) and ability ($\underline{\theta}_{jE}$) parameters are transformed in the following ways:

$$\underline{a}_{iB}^* = k\, \underline{a}_{iE}'\, \mathbf{T}, \tag{10}$$

$$d_{iB}^* = d_{iE} + (\underline{a}_{iE}'\, \mathbf{T})\, \underline{m}\,, \text{ and} \tag{11}$$

$$\underline{\theta}_{jB}^* = (1/k)(\mathbf{T}^{-1}\underline{\theta}_{jB} - \underline{m}), \tag{12}$$

where the superscript * represents the transformed values from the equated scale to the base scale, k is a dilation parameter, **T** is an orthogonal rotation matrix and $\underline{m}$ is a translation vector. In contrast, the following transformations were found in Oshima et al.' study (1997):

$$\underline{a}_{iB}^* = (\mathbf{A}^{-1})'\, \underline{a}_{iE}', \tag{13}$$

$$d_{iB}^* = d_{iE} - (\underline{a}_{iE}'\, \mathbf{A}^{-1})\, \beta, \text{ and} \tag{14}$$

$$\underline{\theta}_{jB}^* = \mathbf{A}\underline{\theta}_{jE} - \beta, \tag{15}$$

where the rotation matrix **A** adjusts the variances and covariances of the ability dimensions, and the translation vector β re-shift the point of origin on each ability dimension. As pointed out previously, no dilation parameter was found or defined in the Oshima et al's MIRT equating methods. If the rotation matrix produced by Oshima et al's study is a nonorthogonal transformation matrix, the relative positions of all pairs of item discrimination parameters in space in the equated form will be distorted while referring them to the base form's reference system.

D. The Methods of Estimating MIRT Transformation Parameters

1. The Principle of Estimating the Rotation Matrix

The ordinary orthogonal procrustes rotation (Schonemann, 1966) is used for rotating the estimates of item discrimination of the equated test so that the sum of the squared differences

(see Equation 16) between each item's pair of item discrimination estimates (base and equated) is *minimized*.

$$\mathbf{E}_I = \mathbf{A}_E \mathbf{T} - \mathbf{A}_B \qquad (16)$$

where $\mathbf{E}_I$ is the residual matrix, $\mathbf{T}$ is the transformation matrix, and $\mathbf{A}_E$ and $\mathbf{A}_B$ are L x m discrimination-parameter matrixes for the equated group and for the base group, respectively (L represents the number of anchor test items and m represents the number of dimension). The order of estimating the scaling coefficients is arbitrary (refer to Schonemann & Carroll, 1970, Li, 1997). Practically speaking, if rotational indeterminacy is resolved, the reminding issues of origin shift and unit change can be resolved using unidimensional IRT equating principles (e.g. minimizing differences between two test characteristic curves).

2. The Principle of Estimating the Scaling Coefficients

Three sets of methods to estimate the scaling coefficients, m_1, m_2 and k were used for this study. Each is discussed in a section to follow.

(1). Matching Test Characteristic Surfaces (MTCS)

The matching test characteristic surfaces method is the extension of Stocking's and Lord's procedures to MIRT models (refer to Oshima, Davey & Lee, 1997). The MIRT version of the test characteristic function is a surface formed by summing the probabilities of correct responses of common items, known as the expected true score. In the two-dimensional case, the correct linear transformation of scales from the common items inserted in two different tests would produce the same expected true score for examinee j if the scaling coefficients, m_1, m_2 and k, were known. In practice, it is desirable to choose m_1, m_2 and k so that the average squared differences between these surfaces is as small as possible. The function below to be minimized is:

$$F = \frac{1}{N} \sum_{j=1}^{N} [t(\theta_{j1},\theta_{j2}) - t^*(\theta_{j1},\theta_{j2})]^2 \qquad (17)$$

where $t(\theta_{j1}, \theta_{j2})$ and $t^*(\theta_{j1}, \theta_{j2})$ are the expected true scores of an examinee on the set of common items in the base test and in the equated test, respectively, and N is the number of grids points. The arbitrary composite abilities can be put into Equation 17. In the two-dimensional case, the number of the grid points equals n^2 if each ability dimension is divided into n points.

The scaling coefficients of m_1, m_2 and k that minimize the value F can be derived by differentiating Equation 17 with respect to m_1, m_2 and k and by setting these three partial derivative equations equal to zero. The scaling coefficients, m_1, m_2 and k can be obtained by simultaneously solving for the three equations, using the Newton-Raphson procedure.

Figure 2 is an example of this process. The lower test characteristic surface in Figure 2 is assumed to be plotted using the first twenty item parameters from ACT Form-24B (Reckase, 1985). On the other hand, the upper test characteristic surface in Figure 2 was graphed using the same test items as those used in the lower test characteristic surface in Figure 2, but the corresponding item parameter estimates are calibrated from another group (Note that the rotational indeterminacy of those estimated item parameters is assumed to be resolved). As a matter of fact, the upper test characteristic surface was plotted using the estimated d parameters by adding 1 on each known d value and the estimated parameter-a's by multiplying 0.5 on each known parameter-a value. It is desirable to seek the scaling parameters ($m_1 = -1$, $m_2 = -1$ and $k = 2$) so that the average squared differences between two test characteristic surfaces is as minimal as possible.

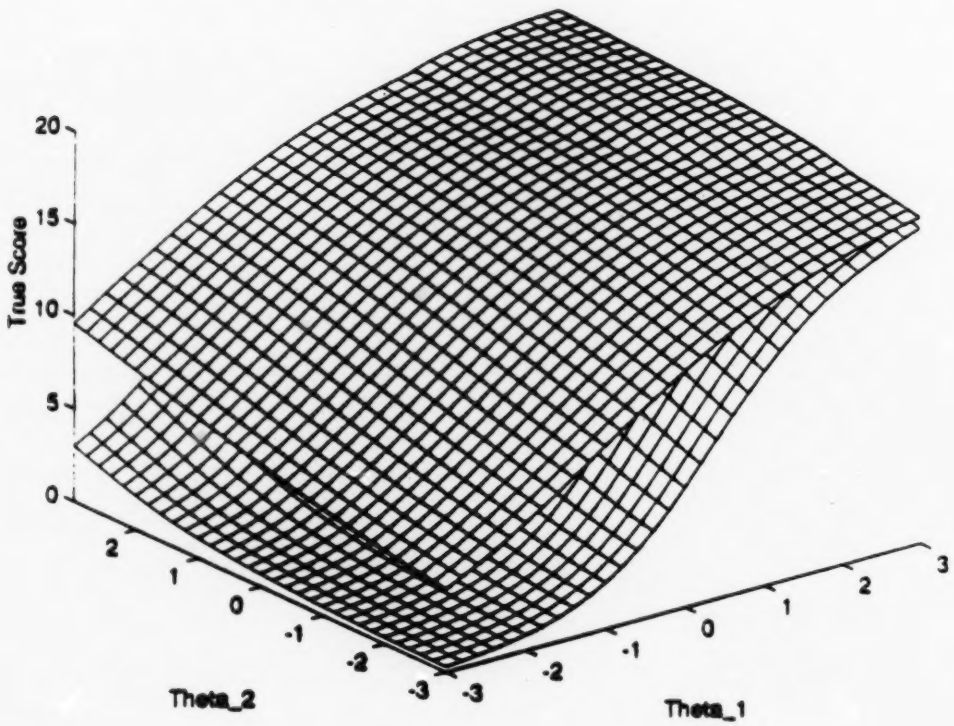


Figure 2. Matching Two Characteristic Surfaces: m_1= -1, m_2=-1, k=2

(2). Least Squares for the translation parameter estimates and Ratio of Eivenvalues for the dilation parameter estimate

The m_1, m_2 and k values can be estimated separately. To find estimators of the m_1 and m_2 parameters, the method of least squares is employed. For each sample observation (d_{iE}, d_{iB}) the method of least squares considers the deviation of d_B from the transformed value of d_{iE} defined

on Equation 11. The method of least squares utilizes the sum of the L (number of common test items) squared deviations. The criterion is denoted by Q:

$$Q = \sum_{i=1}^{L} [\, d_{iB} - d_{iB}^{*} \,]^2 \qquad (18)$$

The values of m_1 and m_2 that minimize Q can be derived by differentiating Equation 18 with respect to m_1 and m_2 and by setting this partial derivative equal to zero. The values of m_1 and m_2 can be obtained by simultaneously solving for the two equations.

The k parameter can be estimated by computing the ratio of the square root of the maximum eigenvalue of **B'B** to the square root of the maximum eigenvalue of **E'E** (called Ratio of Eigenvalues procedure). Let $\mathbf{B} = \mathbf{A'}_B \mathbf{A}_B$, the nonnegative square roots of the eigenvalues of **B'B** are called the singular values of **B**, denoted as Sig (**B**) (see Marcus, 1993). Let $\mathbf{E} = \mathbf{A'}_E \mathbf{A}_E$, the nonnegative square roots of the eigenvalues of **E'E** are called the singular values of **E**, denoted as Sig (**E**). Then the estimate k equals:

$$k = \frac{\text{Max}[\text{Sig}(\mathbf{B})]}{\text{Max}[\text{Sig}(\mathbf{E})]} \qquad (19)$$

where Max represents the Maximum function. In essence, much information about matrix, **B**, can be obtained from the properties of the matrix $\mathbf{B}-\lambda_B\mathbf{I}$, where λ_B values are called the characteristic roots or eigenvalues of **B**. The maximizing values of the λ_B can account for the most variance of the matrix **B** (refer to Tatsuoka & Lohnes, 1988). The square root of the largest eigenvalue of the matrix **B** serves as a reliability index of scaling, denoted as Max[Sig (**B**)]. Similarly, the same information about matrix, **E**, can be obtained. Consequently, the ratio of the Max[Sig (**B**)] to the Max[Sig (**E**)] could be a good estimator of k.

(3). Least Squares for the translation parameter estimates and Ratio of Trace for the dilation parameter estimate

The Least Squares for the translation parameter estimates was illustrated previously. A similar least squares method to estimate the dilation parameter can be found in the study (Schonemann and Carroll (1970). This Least Squares method was developed for fitting one matrix to another under choice of a rotation matrix, a translation vector and a central dilation

vector that minimize the sum of squared errors of the residual matrix $\mathbf{E}_2$ (see Equation 20). The translation transformation is removed from the original formula because the pair of MIRT item discriminations (base test and equated test) can not reveal any information about the translation vector or origin shift coefficients.

$$\mathbf{E}_2 = (k\mathbf{A}_E\mathbf{T}) - \mathbf{A}_B \tag{20}$$

where, k is the unit change coefficient. The above procedure was originally developed for nonmetric multidimensional scaling. This procedure can be applied to MIRT Equating as well. The equations for estimating the rotation matrix and the unit change coefficient were derived simultaneously. After the former step, the estimate of the rotation matrix is free from the estimate of the unit change coefficient (refer to Schonemann & Carroll, 1970). The rotation matrix and the unit change coefficient are obtained through the following steps: (a) Center all the elements of the matrix of $\mathbf{A}_E$, the centered matrix called $\mathbf{A}_{CE}$, (b) Center all the elements of the matrix of $\mathbf{A}_B$, the centered matrix called $\mathbf{A}_{CB}$, (c) Perform a standard orthogonal procrustes subroutine to obtain the transformation matrix $\mathbf{T}$, and (d) Finally, compute the scalar, k

$$k = \frac{\text{trace}(\mathbf{T}'\mathbf{A}'_{CE})\mathbf{A}_{CB}}{\text{trace}(\mathbf{A}'_{CE}\mathbf{A}_{CE})} \tag{21}$$

Based on the above literature review, three sets of MIRT equating procedures could be used for transforming the equated group's reference system into the base group's reference system. The estimates of the transformation parameters can be obtained via the computer program MDEQUATE (Li, 1996), written in the computer language, MATLAB (The MathWorks, Inc, 1995). The numerical differentiation using difference approximations is employed in MDEQUATE to calculate the first-order and the second-order partial derivatives.

BEST COPY AVAILABLE

E. The Distribution of Errors in the Estimation of Parameters

The magnitude of the value of an item parameter itself may have an effect on its standard error. On the average, the hard items and easy items have larger standard errors; as do the high and low discrimination items. When the distribution of abilities is bell-shaped, the standard error of an item difficulty associated with a high discrimination parameter, is lower than the same item difficulty associated with a low discrimination parameter (see Figures 2, 3 and 4, Thissen & Wainer, 1982). Thus, the combination of a set of item parameters for an item should be taken into account when modeling the standard error in the estimation of parameter estimates. The sample size is also a substantive factor on the standard error of parameter estimates. The larger the sample size; the less standard error the parameter estimate has. A mathematical expression for this relation has been developed by Thissen and Wainer (1982) and is illustrated below.

For an item i, the likelihood of the observed responses for N independent examinees is:

$$L = \prod_{j=1}^{N} P_j^{u} (1 - P_j)^{1-u} \qquad (22)$$

where P can be calculated from a M2PL model, u=1 for correct response; u=0 for incorrect response. The loglikelihood of Equation 22 is

$$\log L = \sum_{j=1}^{N} \left[u \log(P_j) + (1 - u) \log(1 - P_j) \right] \qquad (23)$$

The maximum likelihood estimates of each parameter ($\underline{a}_i$, d_i,) are located where the partial derivatives of Equation 23 are zero. For ease of expression, ξ represents the M2PL item parameters ($\underline{a}_i$, d_i,). Given a density of θ (e.g. multivariate Gaussian with iid $N(\mathbf{0}, \mathbf{I})$), for any parameter ξ_s and ξ_t, the negative expected value of the second derivative of the loglikelihood function, Equation, 23, has the form (refer to Thissen, Wainer, 1982),

$$-E\left(\frac{\partial^2 \log L}{\partial \xi_s \partial \xi_t} \right) = N \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left\{ \left(\frac{1}{PQ} \right) \left(\frac{\partial P(\theta)}{\partial \xi_s} \frac{\partial P(\theta)}{\partial \xi_t} \right) \right\} \Phi_2(\theta) d\theta_1 d\theta_2 \qquad (24)$$

where E is the expectation and Q=1-P. Equation 24 requires the derivatives of $P(\theta)$ with respect to its parameters. These derivatives of $P(\theta)$ can be substituted in Equation 24 to give a 3 x 3 (for the M2PL model) information matrix corresponding to the triplet item parameters (d, a_1 and a_2).

The inverse of that information matrix is the asymptotic variance-covariance matrix of the three parameters. The square roots of the diagonal elements of the variance-covariance matrix are the asymptotic standard errors of the parameters.

The numerical approximation of the multiple integral in Equation 24 can be calculated by the two-dimensional Gauss-Hermite quadrature and is presented in Equation 25 in the two-dimensional case.

$$N\sum_{q_2=1}^{q}\sum_{q_1=1}^{q}\left\{\left(\frac{1}{PQ}\right)\left(\frac{\partial P(\underline{X})}{\partial \xi_s}\frac{\partial P(\underline{X})}{\partial \xi_t}\right)\right\}A(X_{q_1})A(X_{q_2}) \qquad (25)$$

where X is a quadrature point in one of two ability dimensions, q is the number of quadrature in this ability dimension and A(X) is the corresponding weight of the quadrature. The number of quadrature points for numerical integration are set to ten for each dimension in this study.

III. Methodology

Two simulation studies have been conducted sequentially. The first examines: The Effect of Error in the Parameter Estimates on the Precision in Estimating the Transformation Parameters. Based on the results of the first study, the best MIRT equating method was chosen for the second study: Linking Multidimensional IRT Parameters onto a Known Target Test Metric: The Item Bank Case.

A. First Study.

1 Key Variables

(1) Sample Sizes and Standard Error

The known item parameters are from ACT Form-24B (Reckase,1985). Each simulated observed item parameter from a set of parameters (d, a_1, a_2) for an item was computed from the summation of the corresponding true item parameter and the expected measurement error, generated as a random value from N (**0**, **V**), where **V** is the asymptotic variance-covariance matrix corresponding to this set of triple item parameters (d, a_1, a_2). Equation 25 was used for computing the matrix **V** for each item under different sample sizes, which are: (a). 1000 (b). 2000, and (c). 4000. The second condition (N=2000) was recommended by several researchers (e.g., Ackerman, 1994; Carlson, 1987) and may serve as a base for comparisons with the rest of the two conditions

(2) Equating Situations

Two cases were explored. They are: (a) Simulation Study of Parameter Recovery or of Equating: Error of the parameter estimates only exists in the equated test. (b) Equating Study of Real Dataset: Error of the parameter estimates exists in both the equated test and the base test.

(3) The Number of Anchor Test Items

Two cases of the number of anchor test items were systematically chosen from Form-24B. They are: (a) 15 anchor test items and (b) 25 anchor test items. Since the order of item number reported in the study (Reckase, 1985) was approximately arranged by the value of traditional difficulty (p value), systematically sampling items from this set of test items will make the test characteristics of the anchor test similar to the ones of the whole test.

(4). Horizontal and Vertical Linking

Two types of item linking were explored. They are: (a). Horizontal Linking: (m_1=0; m_2=0; k=1) (b). Vertical Linking: (m_1= -0.5; m_2=0.5; k=1.25), where m_1 and m_2 are the origin shift coefficients and k is the unit change coefficient. The angle corresponding to the cosine of the rotation was set to 45 degree across all conditions.

2. Data Analysis and Evaluation of Result

In all, there are 72 different combinations of situations (Three methods x Three Sample Sizes x Two Equating Situations x Two Anchor Levels x Two Linking Situations). For each combination, 200 replications have been conducted. The accuracy of the transformation-parameter estimation procedures for each of the 72 simulation conditions was analyzed using two criteria BIAS and RMSE.

Separate regression analyses were performed to predict Log[RMSE] in each of the transformation-parameter estimates (DEGREE, m_1, m_2 and k) by fitting those predictors mentioned above (refer to Harwell, Stone, Hsu and Kirisci, 1996). It should be noted that a log transformation for the outcome variable RMSE was conducted in order to better satisfy (approximate) the normality assumption. In addition, the main reason for choosing a multiple regression method rather than an analysis of variance (ANOVA) as an inferential approach is that some of the simulation factors such as sample size and test length are quantitative.

Since two methods are available to estimate each translation parameter (m_1 or m_2) in this study, an indicator variable (dummy variable) was coded for representing the method of estimating the translation parameter (Least Squares Procedure coded as 1; MTCS coded as 0).

BEST COPY AVAILABLE

Similarly, two equating situations (Horizontal coded as 0; Vertical coded as 1) and two study situations (Parameter Recovery coded as 0; Equating coded as 1) were separately coded as dummy variables. Meanwhile, since three methods are available to estimate the dilation parameter, k, two dummy variables (called as DM01 and DM02) were coded for representing the three methods of estimating k, where the Ratio of Eigenvalue was coded "0 0" as a reference method. The significant t-test for each dummy-variable regression coefficient of "Equating Method" was performed to detect whether one linking method has significantly different impact than the other on the precision of a transformation-parameter estimate when the rest of the simulation factors were held constant.

B. Second Study

1. The Simulation of Test Data

(1) Sample Size and Ability Composites in the Equated Group

The sample size is set to 2000. Two combinations of ability composites in the equated group were generated. They are: (a). Normal distribution on both Dimensions (Called NorNor) (b). Normal distribution on the first-Dimension and positively-skewed distribution on the second-Dimension (Called NorPos). For the tested group of NorNor, two thousand sets of the two-dimension ability parameters were randomly selected from the multivariate normal distribution, N(**0**,**I**). This case meets the default of TESTFACT and serves as a base for comparison.

With respect to the tested group of NorPos, the first-dimension ability parameters were chosen from the first-dimension parameters in the NorNor ability composites. This will make the comparisons of the accuracy of the first-dimension of transformed ability estimates produced from different research conditions possible. Also, the two thousand numerical values randomly selected from a positively-skewed distribution. They were then standardized to a mean of 0 and a SD of 1 and were used as the two thousand ability parameters of the second dimension. This positively-skewed distribution is characterized by a chi-square distribution with eight degrees of

freedom (see Seong, 1990). The skewness of this distribution is 1. All ability parameters were held constant across the 100 replications of data under each combination of study conditions.

(2). Number of Anchor Items and the Whole test

Twenty anchor items were systematically chosen from the 40-item test of Form-24B and used in the equating. The mean MIRT item difficulty, d, of the whole test is -0.324, reflecting a moderately difficult test when ability is centered on a mean of zero mean a SD of one.

(3). Dimensionality

Two kinds of datasets were manipulated. (a) Two-dimensional dataset (b) Unidimensional dataset . A unidimensional dataset was created by setting the item discrimination parameters of the second dimension to zero at the process of data generation.

(4). Data Generation and Parameter Estimation

There exists three research conditions. They are (1) multidimensional NorNor case, (2) multidimensional NorPos case, and (3) unidimensional case. It is assumed that the metric of the target test item parameters is defined with reference to orthogonal traits. The item response vectors were generated via MDGEN01 computer program (Li, 1996) based on the two-dimensional logistic function, M2PL. One hundred replications were generated under each research condition.

The item response datasets were calibrated via TESTFACT, in which several key options were set up as follows: (a) For initial MINRES factor analysis, two-factor solution was requested and the maximum number of iterative communality improvement was set to 5 with the precision criterion for communality improvement at 0.001, (b) For full-information analysis, the number of quadrature points for numerical integration was set to ten for each dimension, the maximum number of E-step iterations and iterations within M-step were set to 100, and the precision criterion for convergence in M-step was set at 0.001, (c) The factor loadings from the full-information item factor analysis were rotated orthogonally according to the "Varimax" criterion

(refer to Muraki & Engelhard, 1985; Bock, Gibbons & Muraki, 1988), and (d) The multidimensional-ability scores (θ) were computed by the expected a posteriori method (EAP).

2. Data Analysis and Evaluation of Result

The effect of the accuracy of the equating method (used in the second study) on linking ability and item parameters onto a common scale is assessed by using BIAS and RMSE criteria. Since the same set of item parameters were repeatedly estimated across research conditions, the Log[RMSE] of each of various item parameter estimates can be treated as a repeated-measure across research conditions. The t-test for dependent observations was then performed to compare the impact of the research condition (e.g. Condition 1 Versus 2 or Condition 1 Versus 3) on the precision of each of various item parameter estimates.

IV. Results and Discussions

A. First Study

1. Descriptive Method: BIAS and RMSE for Each Transformation-parameter Estimate

The combinations of two equating situations (Horizontal and Vertical) and two study situations (Parameter Recovery and Equating) were chosen to generate four main research situations. The descriptive statistics of BIAS and RMSE for each of the MIRT linking coefficients under the first research situation, the combination of horizontal linking and parameter recovery study, are reported in Table 1. Similarly, the descriptive statistics of BIAS and RMSE for the MIRT linking coefficients under the second research situation (the combination of vertical linking with parameter recovery study), the third research situation (the combination of horizontal linking with equating study), and the fourth research situation (the combination of vertical linking with equating study) are reported in Table 1.

The descriptive statistics are sequentially listed under the simulation factors of sample sizes, test length and MIRT equating method. For example, the value of -0.0027, reported in the second data row and first data column of Table 1, represents the BIAS statistic of the k estimate

by the Ratio of Eigenvalue method under the combination of horizontal linking, parameter recovery study, sample size=1000 and test length= 15.

The rotation matrix produced from the procrusts procedure (labeled as Procrus in the tables) is presented as the DEGREE of angle corresponding to the cosine of the rotation. Three methods were used for estimating k. They are labeled as Eig_k (Ratio of Eigenvalues), Trace_k (Ratio of Trace) and MTCS_k (Matching Test Characteristics Surfaces). Two methods were used for estimating m_1 (or m_2). They are labeled as LS_m1 (or LS_m2) and MTCS_m1 (or MTCS_m2) for representing the Least Squares and the MTCS method, respectively. All the BIAS and RMSE indices listed in Table 1 were estimated under the circumstance that the errors in the estimation of item parameters were purposely manipulated.

Findings from Table 1 are summarized: (1). Of these three methods for estimating dilation parameter, k, the Ratio of Trace consistently performed the best in term of the criterion of RMSE and Range (distance between the minimum and the maximum parameter estimates) across all simulation situations; (2). Of the two methods for estimating translation parameters, m_1 and m_2, the Least Squares Procedures consistently performed better across all simulation situations in terms of the criterion of RMSE and Range; (3). The BIAS value produced from each of MIRT estimating methods across all combinations of conditions are close to zero. The results imply that each of these methods for the MIRT equating is a unbiased estimator; (4). The magnitudes of RMSE produced from each of the MIRT equating methods across all simulation conditions are relatively small. It implies that these MIRT equating methods are all effective estimators; and (5). The value of BIAS produced from each of the MIRT estimating methods across all combinations of conditions become small as sample size or test length increases. It implies that these MIRT equating methods are consistent estimators.

Table 1

Descriptive Statistics of BIAS and RMSE of the MIRT Transformation-parameter Estimates under the Four Research Situations

Research Situation	1		2		3		4	
N=1000								
TL=15	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
Procrus	-0.1603	2.2069	-0.0853	2.1947	0.0929	2.3827	-0.0252	2.4481
Eig_k	0.0027	0.0285	-0.0037	0.0384	0.0096	0.0504	0.0013	0.0635
Trace_k	0.0010	0.0259	0.0019	0.0354	0.0024	0.0475	-0.0039	0.0603
MTCS_k	0.0268	0.0731	0.0459	0.1021	-0.0121	0.0834	-0.0041	0.1048
LS_m1	0.0065	0.0679	0.0404	0.1081	-0.0063	0.1025	0.0330	0.1299
MTCS_m1	0.0276	0.0820	0.0485	0.1232	-0.0003	0.1130	0.0225	0.1397
LS_m2	-0.0070	0.0843	-0.0622	0.1338	0.0088	0.1466	-0.0641	0.1791
MTCS_m2	-0.0663	0.1383	-0.1313	0.2232	-0.0127	0.1797	-0.0618	0.2447
TL=25	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
Procrus	-0.0382	1.3897	0.0290	1.5959	-0.0630	1.9569	-0.0885	1.9828
Eig_k	-0.0025	0.0230	-0.0018	0.0316	0.0053	0.0420	0.0059	0.0540
Trace_k	0.0019	0.0212	0.0055	0.0294	0.0016	0.0384	0.0025	0.0498
MTCS_k	0.0348	0.0694	0.0681	0.1119	-0.0021	0.0729	0.0144	0.0971
LS_m1	-0.0024	0.0581	0.0304	0.0914	-0.0094	0.0771	0.0191	0.1078
MTCS_m1	0.0209	0.0699	0.0533	0.1169	-0.0044	0.0904	0.0175	0.1279
LS_m2	-0.0017	0.0579	-0.0401	0.0935	0.0105	0.0819	-0.0325	0.1153
MTCS_m2	-0.0529	0.1021	-0.1156	0.1891	-0.0009	0.1231	-0.0557	0.1787
N=2000								
TL=15	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
Procrus	-0.1227	1.3873	-0.1788	1.3898	0.0280	1.5539	-0.2249	1.7933
Eig_k	-0.0022	0.0173	-0.0030	0.0232	0.0040	0.0394	0.0026	0.0489
Trace_k	0.0005	0.0163	0.0007	0.0224	0.0009	0.0359	-0.0009	0.0457
MTCS_k	0.0147	0.0528	0.0247	0.0706	-0.0124	0.0656	-0.0088	0.0712
LS_m1	0.0011	0.0542	0.0195	0.0801	-0.0016	0.0656	0.0180	0.0871
MTCS_m1	0.0100	0.0556	0.0205	0.0836	-0.0030	0.0700	0.0070	0.0911
LS_m2	-0.0038	0.0709	-0.0378	0.1033	0.0066	0.0878	-0.0357	0.1232
MTCS_m2	-0.0320	0.1011	-0.0667	0.1540	0.0071	0.1190	-0.0302	0.1528
TL=25	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	M BIAS	RMSE
Procrus	0.0280	1.1211	-0.0052	1.0610	-0.0813	1.3537	-0.0853	1.3657
Eig_k	-0.0044	0.0183	-0.0054	0.0230	0.0001	0.0273	-0.0004	0.0353
Trace_k	-0.0013	0.0158	-0.0016	0.0210	-0.0011	0.0248	-0.0014	0.0324
MTCS_k	0.0159	0.0482	0.0263	0.0677	-0.0036	0.0515	0.0034	0.0672
LS_m1	0.0014	0.0438	0.0140	0.0655	-0.0052	0.0587	0.0083	0.0828
MTCS_m1	0.0154	0.0501	0.0257	0.0772	0.0009	0.0625	0.0134	0.0909
LS_m2	-0.0039	0.0439	-0.0242	0.0653	0.0029	0.0596	-0.0192	0.0858
MTCS_m2	-0.0348	0.0756	-0.0626	0.1185	-0.0053	0.0846	-0.0339	0.1260
N=4000								
TL=15	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
Procrus	-0.0888	1.1038	-0.0348	0.9586	0.0470	1.2185	-0.0545	1.2517
Eig_k	0.0000	0.0142	-0.0023	0.0193	0.0031	0.0256	0.0001	0.0339
Trace_k	0.0007	0.0129	-0.0005	0.0180	0.0017	0.0233	-0.0015	0.0316
MTCS_k	0.0069	0.0353	0.0068	0.0456	-0.0007	0.0447	-0.0100	0.0597
LS_m1	0.0032	0.0352	0.0062	0.0510	-0.0024	0.0484	0.0006	0.0655
MTCS_m1	0.0091	0.0394	0.0062	0.0547	-0.0010	0.0514	-0.0002	0.0712
LS_m2	-0.0004	0.0443	-0.0136	0.0638	0.0037	0.0655	-0.0020	0.0903
MTCS_m2	-0.0200	0.0654	-0.0248	0.0953	-0.0031	0.0905	-0.0023	0.1215
TL=25	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
Procrus	0.0192	0.7615	0.0285	0.8149	-0.0528	0.9364	-0.0589	0.9469
Eig_k	-0.0009	0.0111	-0.0014	0.0157	0.0004	0.0201	0.0002	0.0260
Trace_k	-0.0003	0.0098	0.0005	0.0145	-0.0004	0.0186	-0.0005	0.0243
MTCS_k	0.0091	0.0307	0.0141	0.0445	-0.0042	0.0366	-0.0014	0.0475
LS_m1	0.0011	0.0292	0.0083	0.0437	0.0002	0.0405	0.0063	0.0589
MTCS_m1	0.0082	0.0344	0.0149	0.0547	-0.0016	0.0490	0.0010	0.0704
LS_m2	-0.0012	0.0285	-0.0123	0.0412	0.0005	0.0418	-0.0095	0.0581
MTCS_m2	-0.0154	0.0493	-0.0323	0.0811	0.0051	0.0623	-0.0039	0.0895

BEST COPY AVAILABLE

2. Inferential Method: A Regression Approach

Separate regression analyses were performed to predict Log[RMSE] in each of the transformation-parameter estimates (DEGREE, m_1, m_2 and k) by fitting those simulation factors. The adjusted R^2 and standardized regression coefficients in each predictor in the regression model are presented in Table 2.

Regarding the regression model of the Log[RMSE] of k estimate, the predictor of Equating Method, especially for the dummy variable DM2 (MTCS versus Ratio of Eigenvalue), was the main contributor to the variation of the k estimate (see Table 2).The DM2's standardized regression coefficient was 0.652 which was statistically significant from zero. In contrast, the DM1's standardized regression coefficient (Ratio of Trace versus Ratio of Eigenvalue) was -0.067 which was not statistically significant from zero. These results indicate that the performance of these three methods of estimating k was in this order, from best to worst, Ratio of Trace, Ratio of Eigenvalue, and MCTS. The first two equating methods were not statistically different; whereas the comparison between the first two methods and the third one did have statistically significant impact on the precision of the k estimate.

The regression model of Log[RMSE] for the m_1 or m_2 estimates, the standardized regression coefficient of the dummy variable (MTCS versus Least Square Procedures) was statistically different from zero. These results indicate that the Least Squares Procedures did account for less variation of the translation parameter estimate, especially for the m_2 estimate, than the MTCS method did.

The results of adjusted R^2s for each of the Log[RMSE] models, ranging from 0.86 to 0.90, reported on the right side of Table 2, suggest that this set of simulation factors was very sensitive to variations in each of the transformation-parameter estimates. Thus, the accuracy of estimating each of MIRT transformation-parameter estimates appeared to depend heavily on these simulation factors.

Table 2
The Adjusted R^2 and Standardized Regression Coefficients in Each Regression Model Using those Simulation Factors to Predict the Log[RMSE] of the Transformation-parameter Estimate

	Predictors					Adjusted R^2
Outcome	Equating Method	Sample Size	Test Length	Equating Situation	Study Situation	
DEGREE	N/A	-0.837*	-0.395*	0.285*	0.029	0.86*
m1	0.165*	-0.749*	-0.168*	0.315*	0.489*	0.90*
m2	0.452*	-0.622*	-0.378*	0.245*	0.404*	0.90*
	DM01 DM02					
k	-0.067 0.652*	-0.497*	-0.148*	0.361*	0.250*	0.86*
	(Trace vs Eig) (MTCS vs Eig)					

#: For estimating the angle degree of reference system being rotated, only one method was employed so that the predictor of Method was unable to be included in the regression model.

B. Second Study:

1. Item Parameter Recovery:

The plots of BIAS index against the true parameters are used to illustrate the degree of precision in estimating each level (e.g., easy or difficult item; low or high discrimination item, etc.) of item parameter. The descriptive statistics of these two indices of BIAS and RMSE computed across the 40 items provide a summary of global information about the precision in estimating the 40 item parameters.

In the case of normal distribution on both ability dimensions, the plots of BIASs against the true item parameters are presented in Figures 3, 4 and 5 for the item difficulty (d) and the item discriminations (a_1, a_2), in that order. The BIAS statistic for each level of item parameter is usually expected to be close to zero as happened here. For the item difficulty estimates, the average BIAS's and RMSE's of the 40 items were found to be 0.024 and 0.210, respectively. For the item discrimination estimates, the average BIAS's of the 40 items were 0.004 and -0.006 for a_1 and a_2, respectively; the average RMSE amounted to 0.094 and 0.082 for a_1 and a_2, respectively. Since the first research condition meets the default of TESTFACT, the results from this condition would serve as a base for comparisons with those from the second and the third research conditions.

BEST COPY AVAILABLE

Table 3
The Descriptive Statistics of BIAS and RMSE Indices for the Various Item Parameter Estimates and the t-test for Each Repeated Measure of Log[RMSE] under the Comparison of Condition 1 versus Condition 2 (or 3) (Items =40, N=2000)

Condition		BIAS			RMSE			
		Mean	Min	Max	Mean	t	Min	Max
1	d	0.024	-0.73	0.46	0.210		0.04	0.73
	a1	0.004	-0.14	0.11	0.094		0.04	0.17
	a2	-0.006	-0.09	0.08	0.082		0.04	0.18
2	d	0.026	-0.69	0.47	0.205	0.87	0.04	0.70
	a1	0.005	-0.09	0.07	0.079	6.07*	0.04	0.13
	a2	0.003	-0.08	0.08	0.077	2.82*	0.04	0.20
3	d	0.122	-0.63	0.72	0.291	-3.08*	0.09	0.73
	a1	0.005	-0.05	0.03	0.059	7.93*	0.03	0.11
	a2	0.001	-0.03	0.03	0.123	N/A	0.09	0.17

1. MIRT Case: Normal Distribution on Both Dimensions
2. MIRT Case: Normal Distribution on the First Dimension and Positively-skewed Distribution on the Second Dimension
3. UIRT Case

*. P <0.05

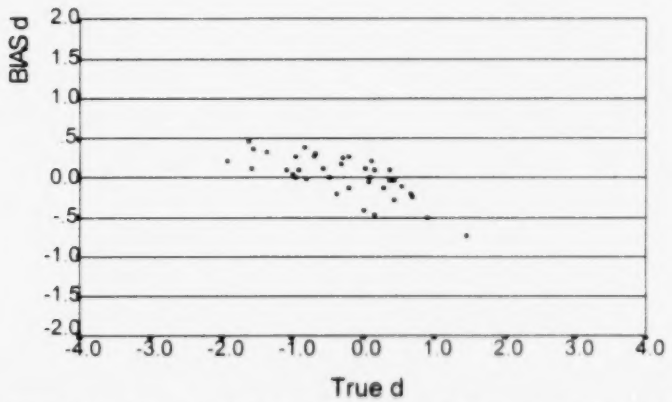


Figure 3. The Plot of BIAS d Parameters Versus the True d Parameters under Condition 1

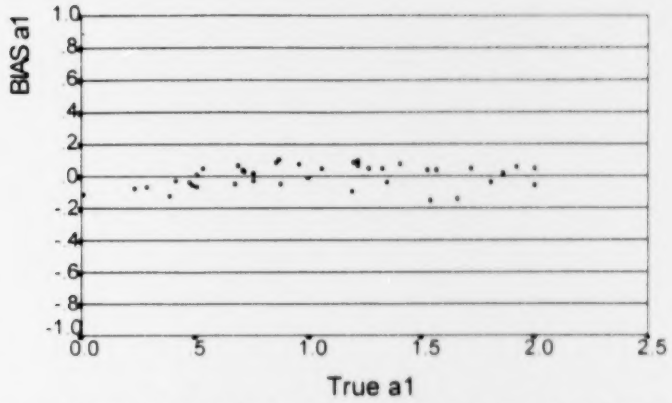


Figure 4. The Plot of BIAS a_1 Parameters Versus the True a_1 Parameters under Condition 1

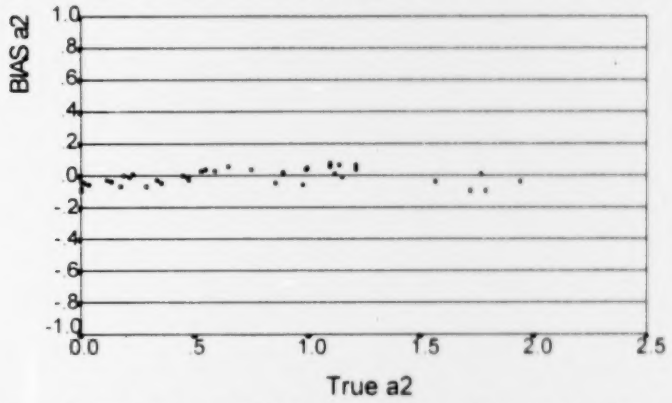


Figure 5. The Plot of BIAS a_2 Parameters Versus the True a_2 Parameters under Condition 1

As regards the case of normal distribution on the first dimension and positively-skewed distribution on the second dimension, the plots in Figures 6 to 8 below present the relationship between the BIAS parameter estimates and the corresponding true item difficulty (d) and discriminations (a_1, a_2) parameters, in that order. These graphical analyses clearly indicate that the results from the first and the second conditions are quite similar. The average BIAS's were 0.026, 0.005 and 0.003 for the item difficulty and item discrimination (a_1, a_2) estimates, respectively (see Table 3). And the average RMSE's were 0.205, 0.079 and 0.077 for the item difficulty and item discriminations (a_1, a_2), respectively. Generally, these average BIAS's and RMSE 's were similar to the corresponding indexes reported on the first research condition.

The dependent t-statistic for the Log[RMSE] of item difficulty estimate indicates that no evidence was found to suggest that the distributions of the multidimensional abilities had a significant impact on the variation of item difficulty estimates. However, a statistically significant difference was found for the RMSE a_1 and RMSE a_2 estimates. As a matter of fact, the RMSE of a_1 and a_2 estimates did decrease rather than increase when the distribution of multidimensional abilities does not meet the default, N (**0,I**) of the TESTFACT.

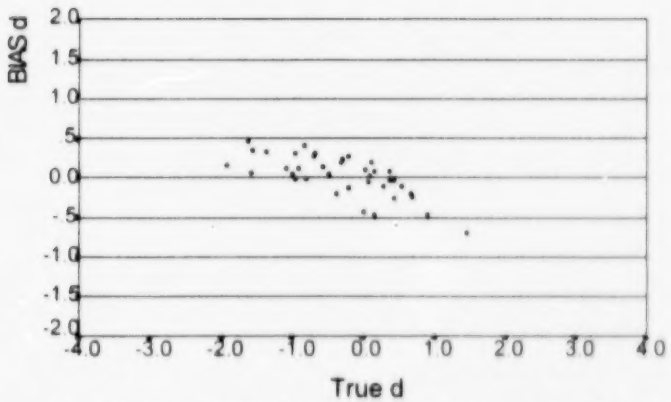


Figure 6. The Plot of BIAS d Parameters Versus the True d Parameters under Condition 2

BEST COPY AVAILABLE

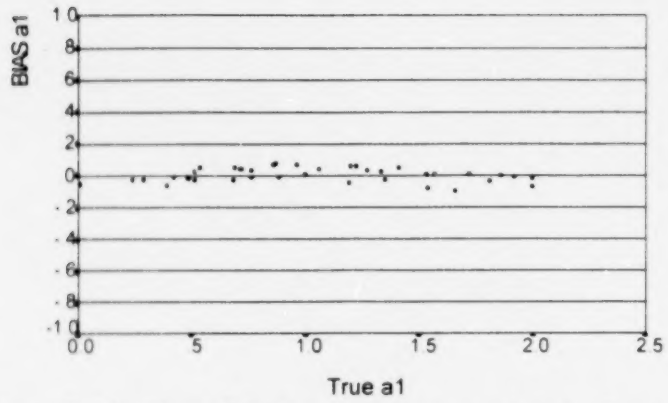


Figure 7. The Plot of BIAS a_1 Parameters against the True a_1 Parameters under Condition 2

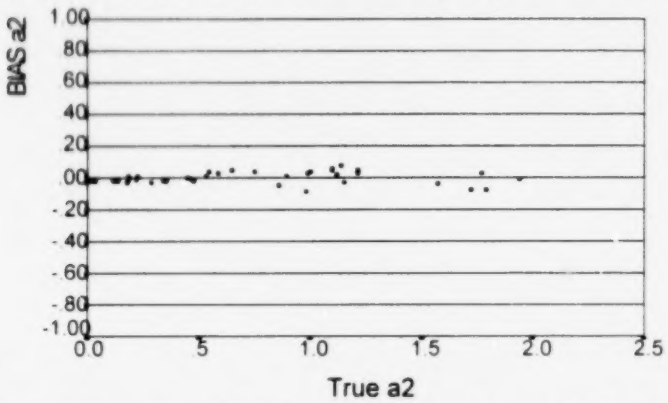


Figure 8. The Plot of BIAS a_2 Parameters Versus the True a_2 Parameters under Condition 2

With respect to the unidimensional case, the plots in Figures 9 and 10 present the relationship between the BIAS statistics and the corresponding true item difficulty (d) and discrimination a_1 parameters, respectively. Comparison between Figures 3 and 9 indicates that although the figures have similar range (-0.7 to 0.7) of BIAS d parameters, a very clear pattern of the BIASs in Figure 9 is found. That is, easy items are always estimated as relatively harder items; hard items are always estimated as relatively easier items and little error is found for the medium difficulty items (d=0). As seen in Figure 10, parameter-a_1 estimates have BIAS that is closer to zero than those seen for parameter-a_1 estimates in Figure 4. Since the true a_2 parameters were all set to zero under the unidimensional case, no plot of BIAS is presented.

The average BIAS was 0.122, 0.005 and 0.001 for the item difficulty, item discrimination (a_1, a_2) estimates, respectively. As for the average RMSE, they amounted to 0.291, 0.059 and 0.123 for the item difficulty, item discrimination (a_1, a_2) estimates, respectively. These BIAS and RMSE indices of item difficulty estimates were slightly larger than the corresponding values obtained from the case of the multidimensional data. In contrast, the RMSE of the a_1 estimates under the case of the unidimensional data was slightly smaller than the corresponding value obtained from the case of the multidimensional data.

The Log[RMSE] of item difficulty estimates produced from the unidimensional data was statistically significantly larger than those produced from the multidimensional data (see Table 3). In contrast, a statistically significant decrease was found for the Log[RMSE] of item discrimination a_1. No comparison of the second-dimension discriminations across conditions was possible. These analyses indicate that more measurement error of the difficulty parameter estimate and less measurement error of the discrimination parameter estimate is likely to occur when the same set of numerical values of item difficulty and discrimination parameters was recovered from the unidimensional case rather than the multidimensional case. From a practitioner's view, based on graphical analyses, the item difficulty and 1st discrimination value from the multidimensional and the unidimensional datasets are quite similar. Accordingly, although the MIRT model is designed to model the multidimensional test data, it is also capable of modeling unidimensional test data.

As pointed out in the section on methodology, the unidimensional dataset was created by setting the second-dimension item discriminations to zero at the time of data generation. This is also necessary while the MIRT equating procedure is used to recover unidimensional IRT item parameters. This MIRT item linking is analogous to linking two-dimensional item discriminations calibrated from both test datasets, in which the estimates of item discriminations on one dimension from two datasets are all close to zero, and may not exist. That is why the MIRT equating procedure can be used successfully to recover unidimensional IRT item parameters.

On the other hand, the same unidimensional test data can also be generated by setting the second-dimension ability rather than discrimination parameters to zero. This setting will cause problems in recovering the item parameter estimates because the number of dimensions of metric is "two" for the true parameters and is "one" for the estimated parameters. We might resolve this problem by setting all the true discrimination parameters of the second dimension to zero during the equating process although they are not really zero. Conceptually, this item linking is analogous to linking one set of the identified item discriminations calibrated from both test datasets, in which one dataset fits a one-dimensional latent trait model well, whereas the other fits a two-dimensional latent trait model well. Consequently, although the results produced from the case of unidimensional datasets indicate that the unidimensional test data can be captured by MIRT model as well, the results also imply that linking a set of identified-unidimensional discriminations of interest in two datasets seems feasible when number of dimensions of the two test datasets is different.

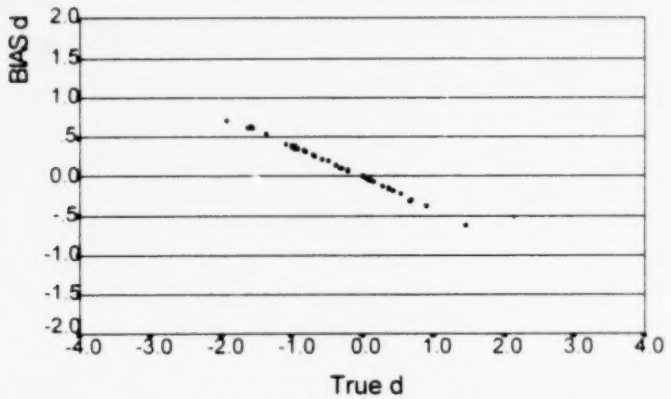


Figure 9. The Plot of BIAS d Parameters Versus the True d Parameters under Condition 3

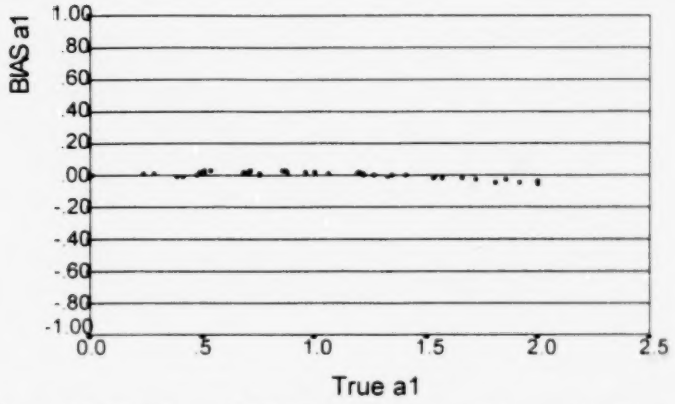


Figure 10. The Plot of BIAS a_1 Parameters Versus the True a_1 Parameters under Condition 3

2. Ability Parameter Recovery

The plots of BIAS statistics against the true parameters are used to illustrate the degree of precision in estimating each level (e.g., low, medium or high) of ability parameter or the expected true score. The descriptive statistics of these two indices of BIAS and RMSE computed across the 2000 examinees provide a summary of global information about the precision in estimating each multidimensional ability estimate and the expected true score.

With regard to the case of normal distribution on both ability dimensions, the plots of BIAS against the true first-dimension and true second-dimension ability parameters are presented in Figures 11 and 12, respectively. The average BIAS's of the 2000 examinees were found to be -0.132 for the first-dimension and 0.355 for the second-dimension ability estimate. Their other corresponding statistics such as standard deviation, skewness, kurtosis, minimum and maximum are also reported in Table 4. Likewise, the average RMSE was 0.336, 0.515 for the first-dimension and the second-dimension ability estimate, respectively. Other corresponding statistics such as minimum and maximum are also reported in Table 4. The descriptive statistics of BIAS and RMSE and graphical analyses indicate that the ability parameters are precisely estimated. The first-dimension discrimination parameters of the simulated test used in this study had, on the average, higher values than the second-dimension ones so that the larger test information of the first-dimension ability estimates can be expected. Consequently, their corresponding standard errors of ability estimates are relatively smaller.

Two issues derived from the above results deserve more attention. One is that the average first-dimension ability estimate was underestimated and the average second-dimension ability estimate was overestimated. The reasons for this result need to be investigated. The other issue is that the low ability parameter was often overestimated and the high ability parameter was often underestimated according to the graphical analysis. This is partly due to the regression effect when the EAP was used to estimate ability parameters.

Table 4
The Descriptive Statistics of the BIASs or the RMSEs Index for the Multidimensional Ability and the Expected True Score Estimates under Three Simulation Conditions[1,2,3] (N=2000)

Condition		BIAS					RMSE		
	Mean	SD	Skew	Kurt	Min	Max	Mean	Min	Max
1	Nθ1	-0.132	0.257	-0.52	7.64	-2.15	1.49	0.336	0.17 2.16
	Nθ2	0.355	0.328	0.55	3.93	-1.70	2.16	0.515	0.22 2.17
	t	1.38	2.273	0.35	1.14	-6.85	9.86	2.840	0.64 9.98
2	Nθ1	-0.136	0.243	-0.39	10.17	-2.21	1.43	0.337	0.17 2.21
	Pθ2	0.322	0.345	-4.17	46.64	-5.32	1.31	0.516	0.21 5.33
	t	0.835	0.598	-1.05	4.64	-3.30	2.83	1.969	1.06 3.50
3	Nθ1	-0.002	0.133	0.64	23.38	-1.36	1.21	0.255	0.14 1.38
	Nθ2	0.041	1.016	-0.20	-0.09	-3.90	2.95	1.236	0.69 4.00
	t	1.016	0.461	0.03	0.35	-1.33	2.89	2.020	0.94 2.94

Note:
N represents the normal distribution
t represents the expected true score
P represents the positively-skewed distribution

Regarding the expected true score estimates, the graphical analysis plotted the BIASs against the corresponding expected true scores (Figures 13) denotes that the estimate of the expected true score seems somewhat inaccurate. For instance, examinees with extreme expected true scores often get high BIAS scores, ranging from -6.85 to 9.86. The average BIAS's was 1.38, where the expected true score scale ranged from 0 to 40. Likewise, the average RMSE's was 2.84. The expected true score recovery suggests that if test items are designed to measure two latent traits that are uncorrelated with each other and of equivalent relative ability level, the estimated test score may be inaccurate and unreliable. It implies that using the estimated expected true score to reflect examinees' multiple-traits abilities is risky.

BEST COPY AVAILABLE

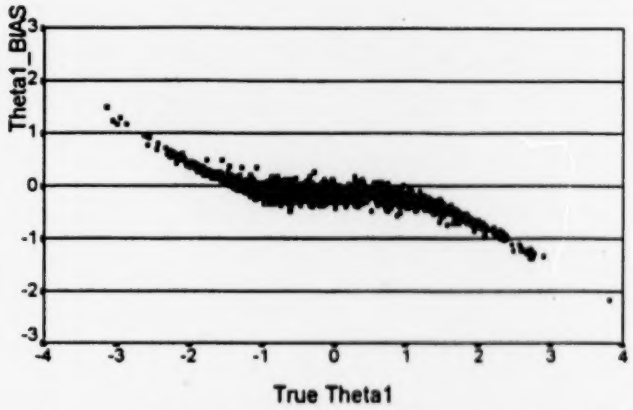


Figure 11. The Plot of BIAS Theta_1 Versus the True Ability Parameters under Condition 1

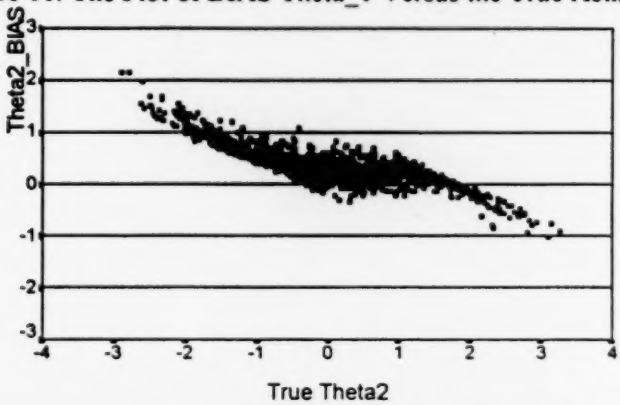


Figure 12. The Plot of BIAS Theta_2 Versus the True Ability Parameters under Condition 1

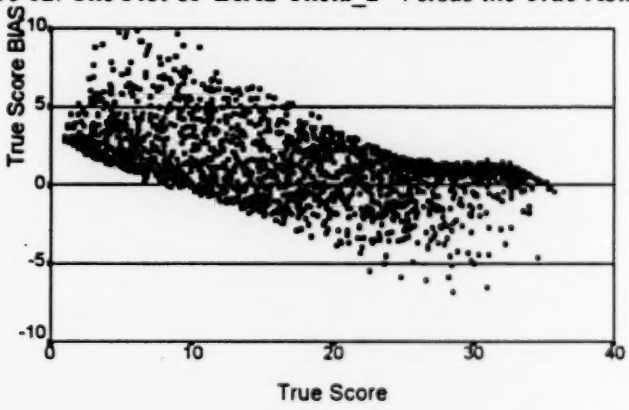


Figure 13. The Plot of BIAS Expected True Scores Versus the Expected True Scores under Condition 1

As regards the case of normal distribution on the first-dimension abilities and positively-skewed distribution on the second-dimension abilities, the plots of BIAS against the true first-dimension and second-dimension ability parameters are presented in Figures 14 and 15, respectively. The average BIAS's of the 2000 examinees were found to be -0.136, 0.322 for the first-dimension and the second-dimension ability estimates, respectively. Their other corresponding statistics such as skewness, kurtosis, minimum and maximum are reported in Table 4. Likewise, the average RMSE were 0.337, 0.516 for the first-dimension and the second-dimension ability estimates, respectively. Their corresponding statistics such as minimum and maximum are also reported in Table 4. Based on the descriptive data and the graphical analyses, the estimates of multidimensional ability parameters were quite precise.

Regarding the expected true score estimates, the graphical analysis plotting the BIASs against the corresponding expected true scores (see Figure 16) denotes that the precision of the expected true score estimate seems more satisfactory than those results from the first simulation condition (Normal Distribution on Both Ability Dimensions). As seen in Table 4, the average BIAS's and RMSE's were 0.835 and 1.97, respectively. These findings suggest that if test items are sensitive to two traits, one of which is minor for most examinees, a single test score report such as the expected true score may be appropriate. This is basically consistent with the observation by Ackerman (1992) that under same circumstances test data can be modeled unidimensionally.

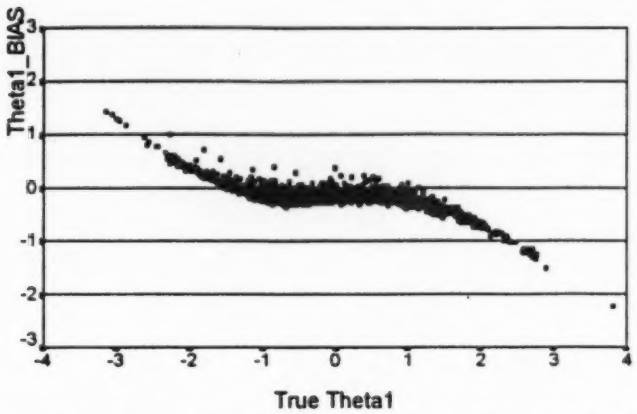


Figure 14. The Plot of BIAS Theta_1 Versus the True Ability Parameters under Condition 2

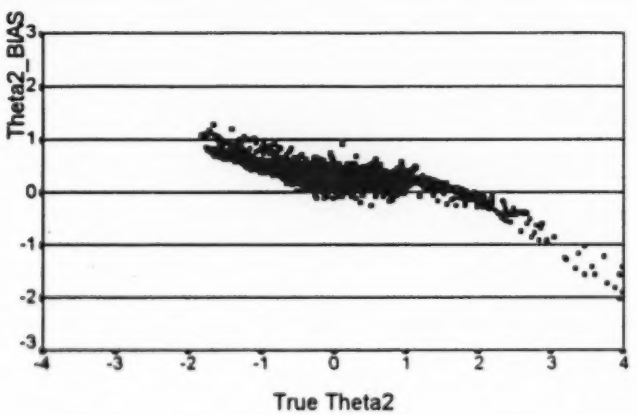


Figure 15. The Plot of BIAS Theta_2 Versus the True Ability Parameters under Condition 2

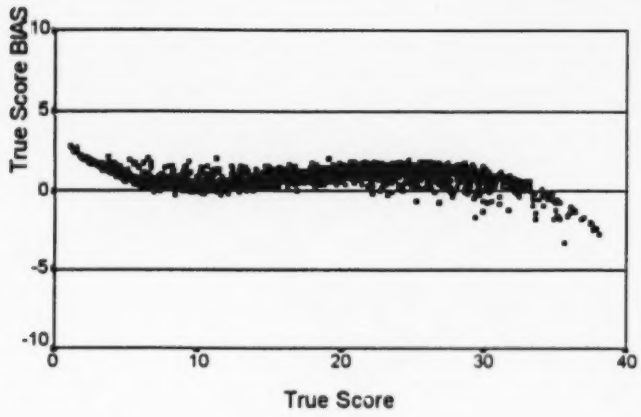


Figure 16. The Plot of BIAS Expected True Scores Versus the Expected True Scores under Condition 2

With respect to the unidimensional case, plotting the BIAS versus the corresponding true ability parameters (see Figure 17) clearly indicates that the first-dimension ability parameters were quite precisely estimated. Comparisons of these two plots with the corresponding Figure 11 (a plot, similar to Figure 17, was presented in the first research condition) and Figure 14 (a plot, similar to Figure 17, was presented in the second research condition) strongly suggest that the ability estimates are potentially more accurate and reliable when the test items are purposely constructed to measure only one trait rather than to measure two traits. The average BIAS's and RMSE's of the 2000 examinees' ability estimates on the first dimension were found to be -0.002 and 0.255, respectively. These two BIAS and RMSE indices from the unidimensional case were smaller than the corresponding values of the BIAS and RMSE for the first dimension ability estimate found in the first and the second research conditions.

The findings have two implications for the field. First, if only one trait is to be measured, the test items should be designed to be sensitive to that latent trait only. Doing so will ensure that the unidimensional ability estimates are more accurate and stable. Second, from another perspective, the test equating procedure for recovering a set of unidimensional ability parameters can be analogous to equating one set of identified-dimension abilities calibrated from two datasets, in which one fits unidimensional trait model well; the other fits a two-dimensional latent trait model well. The success of recovering a set of the identified-dimension ability parameters

indicates that when the number of dimensions of the two test datasets is different, equating one of the identified ability dimensions of interest in two datasets seems feasible. In the same manner, this sort of problem can occur in real testing situations. For example, when two different tests with common items are designed to measure two latent traits of interest and are administered to two different groups, one group of examinees may vary significantly on one of the requisite traits while the other group of examinees varies on both requisite traits. In this case, the dimensionality of the two test response datasets can be quite different, permitting equating on just one dimension.

Within the framework of real testing, both the dimensionality of the target responses and the dimensionality of the responses to be transformed are unknown and are obtained by estimation. The process of identifying the dimensions for each dataset and matching the pairs of the identified dimensions of interest from both test datasets is a critical step in the MIRT equating (refer to Davey, Oshima & Lee, 1996). Put another way, whether the MIRT equating can be successfully accomplished depends on how well the MIRT can fit the data and how clear the latent structure of the test-examinee interaction in each dataset can be revealed. The ability dimensions are statistical constructs and the interpretation of these dimensions may not be essential in Reckase's argument (1977a) about the definition of multiple ability dimensions, when the main concern is the data-model fit.

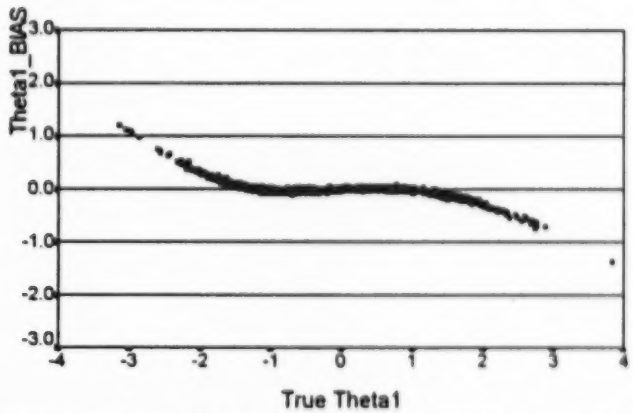


Figure 17. The Plot of BIAS Theta_1 Versus the True Ability Parameters under Condition 3

V. Summary and Conclusions

The three research questions have been explored and the summary of research results are given below. Finally, future research questions resulting from this study is highlighted.

A. Summary and Conclusions to the Research Questions

Regarding the question of " Which MIRT equating method among the three MIRT equating methods developed can produce the most accurate transformation-parameter estimates?", the most appropriate one is the combination of procrustes rotational approach , the Ratio of Trace and the Least Squares Procedures for estimating the rotational transformation matrix, the dilation parameter and the set of translation parameters, respectively.

More specifically, the performance of the procrustes rotational method for estimating the rotation matrix was quite satisfactory in terms of the BIAS and RMSE indices. The ability of these three methods to estimate the dilation parameter is in this order, from best to worst, Ratio of Trace, Ratio of Eigenvalue, and MCTS. In estimating each of the set of translation parameters, the Least Squares Procedures can produce less errors, especially for the m_2 estimate, than the MTCS method did.

With respect to the question of "How accurate can the best MIRT equating method transform parameters onto a target metric?", this MIRT equating method is capable of producing accurate linking of items and "approximately" equivalent estimation of ability parameters under well-fitting model conditions. However, the precision of the expected true score recovery is dependent on the multidimensional-ability composites.

In the case of MIRT ability recovery, one unusual pattern was found and needs further study. That is, the average first-dimension ability estimate was underestimated and the average second-dimension ability estimate was overestimated. With respect to the expected true score recovery, the estimate of the true score was not accurate, especially in the case where the two latent traits are uncorrelated with each other and of equivalent relative ability level. This finding implies that the expected true score can not be a good score to report because the variation of the expected true score estimates is dependent on the composites of multiple-trait parameters as pointed out by (Ackerman, 1996). Luecht (1996) also stressed that if the reported test score is

strictly limited to a single score, there may be little apparent practical advantage in considering the MIRT model that attempts to capture the salient multidimensionality of the test-examinee interaction. Unfortunately, so far, providing the meaning for each of the complete latent traits is difficult and people are not ready to accept the multiple scores generated from a single test. Therefore, the issue of what composite of the latent traits is best measured by a single reported score is critical and has being investigated (e.g., see Zhang , 1997).

Regarding the question of "Can the MIRT model also be applied to a unidimensional test dataset ?", the MIRT model can be applied to unidimensional test data as well. This result also implies that equating one of the identified ability dimensions of interest in two datasets seems feasible when the number of dimensions of the two test datasets is different.

The measurement errors of a_1-parameter estimates in the unidimensional datasets are much smaller than the corresponding a_1-parameter estimates in the two-dimension datasets. The d-parameters are slightly biased. Comparison of the results of ability recovery from the unidimensional test data with those first-dimension ability estimates under the two types of multidimensional test data strongly suggests that the unidimensional ability estimates are more accurate and stable when test items are only sensitive to one latent trait being measured. From another perspective, the test equating procedure for recovering a set of unidimensional ability parameters can be analogous to equating one set of identified-dimension abilities calibrated from two datasets, in which one fits a unidimensional trait model well; the other fits a two-dimensional latent trait model well. The success of recovering a set of the identified-dimension ability parameters indicates that equating one of the identified ability dimensions of interest in two datasets seems feasible when number of dimensions of the two test datasets is different. However, whether equating some identified ability dimensions of interest can be successfully accomplished depends on how well the MIRT can fit the data and how clear the latent structure of the test-examinee interaction in each dataset can be revealed.

B. Future Research Questions

Further research is needed in many related areas. As seen in Table 4, on the average, ability estimates on both dimensions are slightly biased. Example questions related to these results

BEST COPY AVAILABLE

are, Is the sample size too small?, Is the whole test or anchor test too short? Is the parameter estimation method or the corresponding estimation computer program questionable? etc. These, and similar problems have been largely addressed in the area of unidimensional IRT. In contrast, the research on these questions in the multidimensional case is quite rare, especially when the MIRT model increases in complexity to include more dimensions, say Ten (see Thompson, Nering & Davey, 1997).

Although Reckase and Hirsch (1991) suggest that " the number of dimensions is often underestimated and that overestimating the number of dimensions does little harm." We might expect that the suggestion they made is more true if the number of test items and sample size are "very large." Unfortunately, the number of test items is often limited, say Fifty, in real testing situations. The issue raised here is " How many dimensions can be clearly identified by a 50-item test?" If the dimensions are defined as statistical constructs and the interpretation of the dimensions is not essential, this issue may be lessened because the estimation algorithms might allow each item to have multiple high loadings on many dimensions. However, within the framework of MIRT equating, the process of identifying meaningful dimensions is critical, and the goal of interpretable dimensionality may be assisted by instructing the MIRT estimation procedure to produce the simple structure of item factor loadings. That is, an item should only make a high contribution to one dimension and have little contribution to the rest of the dimensions. The combination of simple structure and the limitation of test length may limit the number of dimensions clearly identified. Consequently, test practitioners want to know under which conditions such as test length, number of dimension and sample size the MIRT equating can produce an accurate result.

In this study, we also found that the accuracy of ability estimates in each dimension relies on the corresponding magnitude of the discrimination parameters. If a complex model, say 20 dimensions, is chosen to account for a 50-item test and the simple structure of item discriminations is pursued, we might expect that the values of the discrimination estimates of some dimensions will be too small. As found in this study, ability estimates on the dimension with lower discrimination parameters can be relatively inaccurate or unstable. It may imply that the precision of estimating ability parameters for the dimensions with low item discriminations will be

questionable. Consequently, the question of how many dimensions are appropriate to explain any set of test data deserves careful attention.

Another issue is related to the robustness of MIRT equating with heterogeneous groups. If two tests measure the same multiple latent traits and are administered to two heterogeneous groups which differ in the locations and variabilities of their ability distributions, can the MIRT equating method perform well? In particular, if the underlying dimensionality of the two datasets is different, how can the MIRT equating overcome this problem?

Another quite practical research topic is how do various MIRT equating methods compare in their accuracy of linking of items and equivalent estimation of ability parameters. The MIRT equating procedures developed by Oshima et al (1997) and Thompson et al (1997) are different from those introduced in this study. Practitioners are concerned with which MIRT equating method can produce more accurate linking of the item and ability parameters.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. Journal of Educational Measurement, 29, 67-91.

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. Applied Measurement in Education, 4, 255- 278.

Bock, R D., Gibbons, R. & Muraki, E. (1988). Full-information item factor analysis. Applied Psychological Measurement, 12, 261-280.

Carlson, J. E. (1987). Multidimensional item response theory estimation: A computer program. Unpublished manuscript.

Davey, T., Oshima, T. C. & Lee, K. (1996). Linking multidimensional item calibrations. Applied Psychological Measurement, 20, 405-416.

Green, P. E. (1976). Mathematical tools for applied multivariate analysis. New York: Academic Press.

Hambleton, R. K., Jones, R. W. & Rogers, H. J. (1993). Influence of item parameter estimation errors in test development. Journal of Educational Measurement, 30, 143-155.

Hambleton, R. K. & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston: Kluwer Nijhoff Publishing.

Harwell, M. R., Stone, C. A., Hsu, T. , & Kirisci, L. (1996). Monte carlo studies in item response theory. Applied Psychological Measurement, 20, 101-125.

Hirsch, T. M. (1989). Multidimensional equating. Journal of Educational Measurement, 26, 337-349.

Kolen, M. J. & Brennan, R. L. (1995). Test equating: Methods and practices. New York: Springer-Verlag.

Li, Y. H. (1996). MDEQUATE: A computer program to compute the multidimensional IRT equating parameters, Unpublished manuscript.

Li, Y. H. (1996). MDGEN01: A computer program to generate the multidimensional IRT response patterns, Unpublished manuscript.

Lissitz, R. W., Schonemann, P. H. & Lingoes, J. C. (1976). A solution to the weighted procrustes problem in which the transformation is in agreement with the loss function. Psychometrika, 41, 547-550.

Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. Applied Psychological Measurement, 20, 389-404.

Marcus, M. (1993). Matrices and MATLAB: A tutorial. Englewood Cliffs, NJ: Prentice-Hall, Inc.

McKinley, R. L. & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. Applied Psychological Measurement, 9, 49-57.

Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. Journal of Educational Statistics, 11, 3-31.

Muraki, E. & Engelhard, G. (1985). Full-information item factor analysis: Applications of EAP scores. Applied Psychological Measurement, 9, 417-430.

Oshima, T. C. & Davey, T. (1994, April). Evaluation of procedures for linking multidimensional item calibrations. Paper presented at the annual meeting of National Council on Measurement in Education, New Orleans.

Oshima, T. C., Davey, T. & Lee, S. (1997, March). Multidimensional linking: Four practical approaches. Paper presented at the annual meeting of American Educational Research Association, Chicago, IL.

Reckase, M. D. (1985). The difficulty of items that measure more than one ability. Applied Psychological Measurement, 9, 401-412.

Reckase, M. D. (1997a). A linear logistic multidimensional model for dichotomous item response data. In W. J. Linden and R. K. Hambleton (Eds.), Handbook of modern item response theory (pp. 271-286). New York: Springer-Verlag.

Reckase, M. D. (1997b). The past and future of multidimensional item response theory. Applied Psychological Measurement, 21, 25-36.

Reckase, M. D. & Hirsch, T. M. (1991, April). Interpretation of number correct scores when the true number of dimensions assessed by a test is greater than two. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Reckase, M. D., Thompson, T. D., Nering, M. (1997, June). Identifying similar item content clusters on multiple test forms. Paper presented at the annual meeting of the Psychometric Society, Gatlinburg, TN.

Schonemann, P. H. (1966). A generalized solution of the orthogonal procrustes problem. Psychometrika, 31, 1-10.

Schonemann, P. H. & Carroll, R. M. (1970). Fitting one matrix to another under choice of a central dilation and a rigid motion. Psychometrika, 35, 245-255.

Seong, T. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. Applied Psychological Measurement, 14, 1990.

Skaggs, G. & Lissitz, R. W. (1988). Effect of examinee ability on test equating invariance. Applied Psychological Measurement, 12, 69-82.

Stocking, M. L. (1990). Specifying optimum examinees for item parameter estimation in item response theory. Psychometrika, 3, 461-475.

Tam, H. P. & Li, Y. H. (1997, March). Is the use of the difference likelihood ratio chi-square statistic for comparing nested IRT models justifiable? Paper presented at the annual meeting of American Educational Research Association, Chicago, IL.

Tatsuoka, M. M. & Lohnes, P. R. (1988). Multivariate analysis. New York: Macmillan Publishing Company.

The MathWorks, Inc. (1995). MATLAB: The ultimate computing environment for technical education. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Thissen, D. & Wainer, H. (1982). Some standard errors in item response theory. Psychometrika, 47, 397-412.